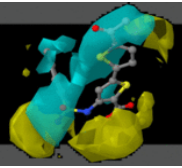
Molecular Similarity and Chemical Similarity

SAPIENZA
UNIVERSITÀ DI ROMA



Overview

Chemical and Molecular Similarity

Molecular similarity importance

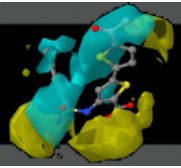
Representations of chemical structures

Similarity and Dissimilarity measures

Similarity searching

Cluster and diversity analysis

What does similarity stand for?



Chemical vs Molecular Similarity

The terms chemical and molecular similarity are often used synonymously but this may not be entirely accurate.

Chemical similarity is based primarily on the physicochemical characteristics of compounds (e.g., solubility, boiling point, log P, molecular weight, electron densities, dipole moments, etc.)

Molecular similarity focuses primarily on the structural features (e.g., shared substructures, ring systems, topologies, etc.) of compounds and their representation.

Similarity perception

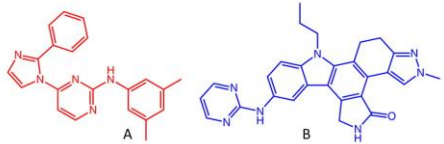
Two *Vascular Endothelial Growth Factor Receptor 2* ligands and different ways to assess their similarity.

Chemical similarity

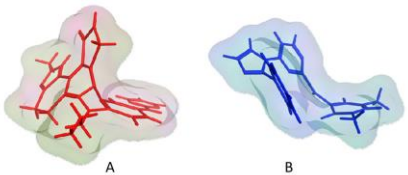
	Mol. weight	LogP	Rotatable bonds	Aromatic rings	Heavy atoms
A	341.4	5.23	4	4	26
B	463.5	4.43	4	5	35

Molecular similarity

2D similarity



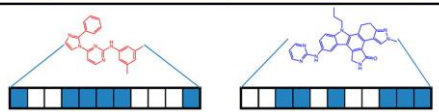
3D similarity



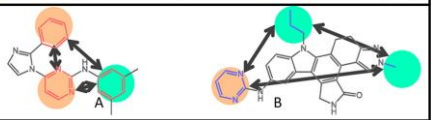
Biological similarity

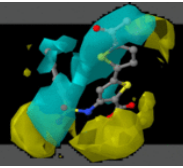
	Vascular endothelial growth factor receptor 2	Tyrosine-protein kinase TIE-2
A	active	inactive
B	active	active

Global similarity



Local similarity



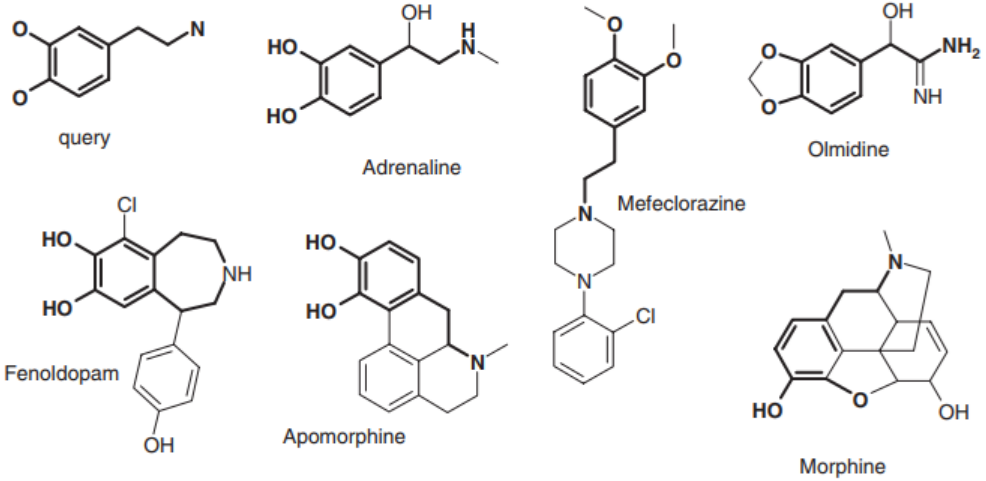


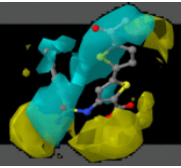
Importance of chemical similarity

The notion of chemical similarity (or molecular similarity) is one of the most important concepts in chemoinformatics.

It plays an important role in predicting the properties of chemical compounds, designing chemicals with a predefined set of properties and, especially, in conducting drug design studies by screening large databases containing structures of available (or potentially available) chemicals.

There are many ways to measure the similarity


query
Adrenaline
Mefeclorazine
Olmidine
Fenoldopam
Apomorphine
Morphine

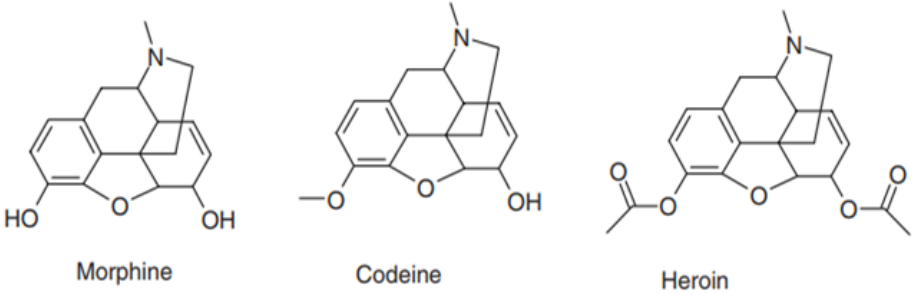


Neighbourhood principle

The rationale for similarity searching lies in the similar property principle [Johnson and Maggiora 1990] which states that structurally similar molecules tend to have similar properties.

Given a molecule of known biological activity, compounds that are structurally similar to it are likely to exhibit the same activity. This characteristic has been referred to as **neighbourhood behaviour** [Patterson et al. 1996].

Neighbourhood principle (2)



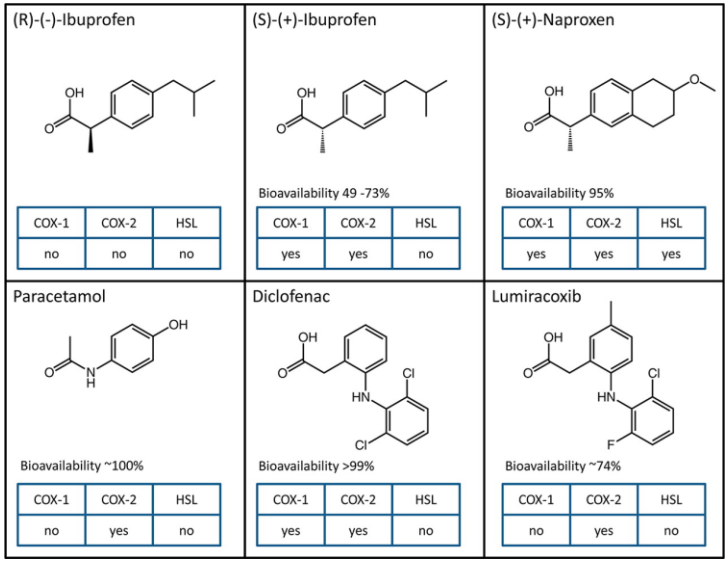
Compounds active at opioid receptors

Neighbourhood principle (3)

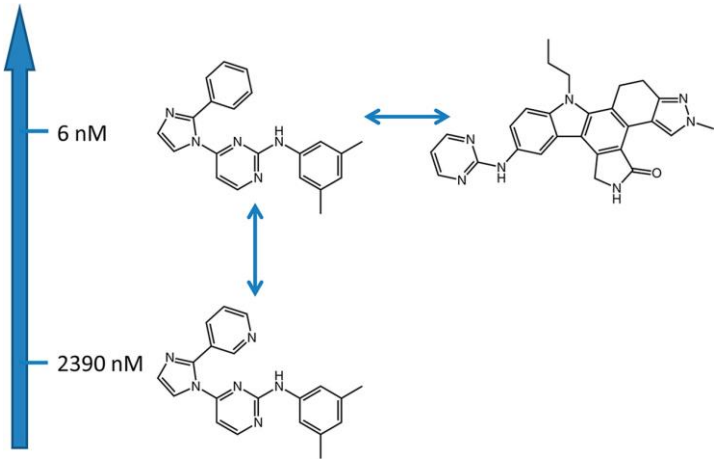
Compounds such as the ibuprofen enantiomers, ibuprofen and paracetamol, or diclofenac and lumiracoxib, appear visibly similar.

From a medicinal chemistry point of view, however, this assessment may not be generally agreed upon since small chemical differences can lead to important changes in specificity profiles (e.g., diclofenac vs lumiracoxib) or compounds containing different functional groups can be synthesized or derivatized in different ways (e.g., ibuprofen vs paracetamol).

Moreover, these COX inhibitors are involved in highly complex similarity–activity relationships (similar mechanism of action) that also cannot easily be separated from a medicinal chemistry perspective.



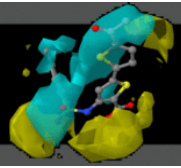
Cyclooxygenase (COX) inhibitors and their activity profiles are compared.
HSL = Hormone-Sensitive Lipase



Similarity versus activity. Three vascular endothelial growth factor receptor 2 ligands are shown that represent different (vertical vs horizontal) similarity–activity (potency) relationships.

There are many exceptions to the principle but it is an excellent rule-of-thumb in the absence of more detailed knowledge

People's judgements of similarity are inherently subjective, so need to provide a quantitative basis, a similarity measure, for assessing the degree of resemblance

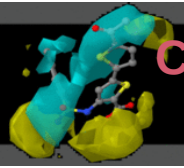


Components of a similarity measure

The main difficulty with similarity searching is that assessing the degree of similarity between two objects is subjective.

In order to be able to quantify the similarity between two molecules, a similarity searching method requires two components:

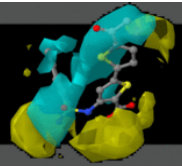
- **a set of numerical descriptors** that can be used to compare molecules;
- **a similarity coefficient** which provides a way of quantifying the degree of similarity based on the descriptors.



Computer representations of chemical structures

Numerical values assigned to structures:

- 1D properties: MW, logP, PSA etc
- 2D properties: fingerprints, topological indices
- Maximum Common Substructures
- 3D properties: molecular fields, shape



Molecular descriptors

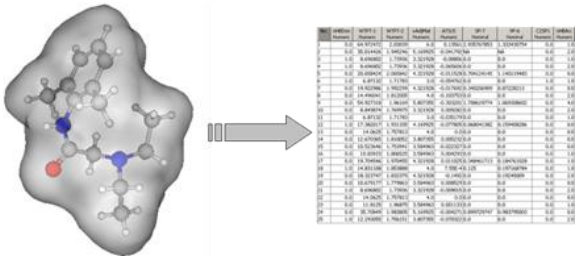
0D: bond counts, molecular weight, atom counts

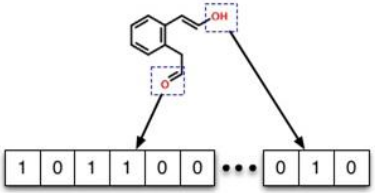
1D: fragment counts, H-Bond acc/don, Crippen, PSA, SMARTS

2D: topological descriptors (Balaban, Randic, Wiener, BCUT, kappa, chi)

3D: geometrical descriptors, surface properties, COMFA

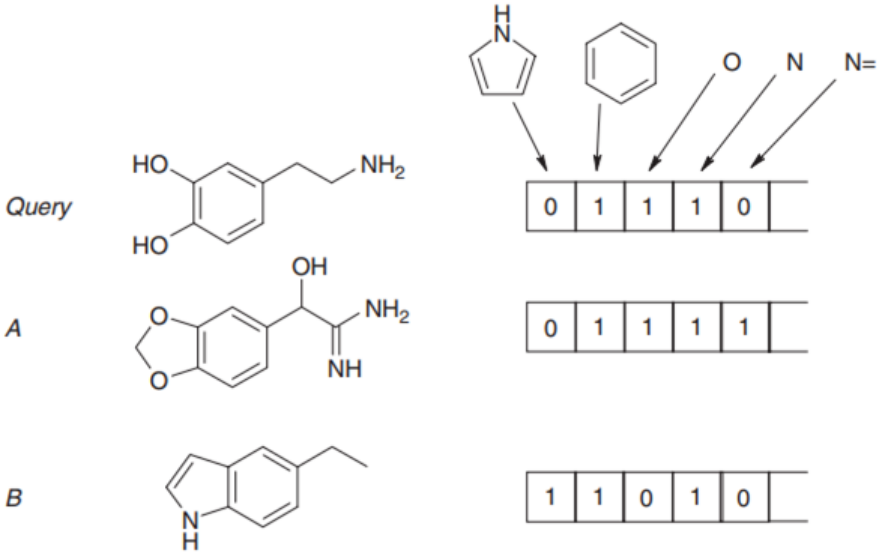
4D: 3D coordinates + conformations

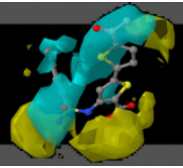




- Lots of types of fingerprints
- "Keyed" fingerprints indicate the presence or absence of a structural feature
- Length can vary from 166 to 4096 bits or more
- Fingerprints usually compared using the Tanimoto metric

Fingerprints: an example





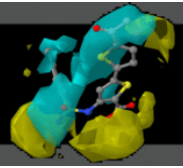
Similarity measures

by www.RCMD.it

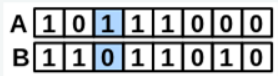
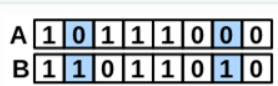
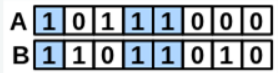
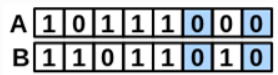
Name	Formula for continuous variables	Formula for binary (dichotomous) variables
Tanimoto (Jaccard) coefficient)	$S_{AB} = \frac{\sum_{i=1}^{N} x_{iA}x_{iB}}{\sum_{i=1}^{N}(x_{iA})^2+\sum_{i=1}^{N}(x_{iB})^2-\sum_{i=1}^{N}x_{iA}x_{iB}}$ Range: -0.333 to $+1$	$S_{AB} = \frac{c}{a+b-c}$ Range: 0 to 1
Dice coefficient (Hodgkin index)	$S_{AB} = \frac{2\sum_{i=1}^{N} x_{iA}x_{iB}}{\sum_{i=1}^{N}(x_{iA})^2+\sum_{i=1}^{N}(x_{iB})^2}$ Range: -1 to $+1$	$S_{AB} = \frac{2c}{a+b}$ Range: 0 to 1
Cosine similarity (Carbó index)	$S_{AB} = \frac{\sum_{i=1}^{N} x_{iA}x_{iB}}{\left[\sum_{i=1}^{N}(x_{iA})^2 \sum_{i=1}^{N}(x_{iB})^2\right]^{1/2}}$ Range: -1 to $+1$	$S_{AB} = \frac{c}{\sqrt{ab}}$ Range: 0 to 1
Euclidean distance	$D_{AB} = \left[\sum_{i=1}^{N}(x_{iA}-x_{iB})^2\right]^{1/2}$ Range: 0 to ∞	$D_{AB} = \sqrt{a+b-2c}$ Range: 0 to N
Hamming (Manhattan or City-block) distance	$D_{AB} = \sum_{i=1}^{N}\lvert x_{iA}-x_{iB}\rvert$ Range: 0 to ∞	$D_{AB} = a+b-2c$ Range: 0 to N
Soergel distance	$D_{AB} = \frac{\sum_{i=1}^{N}\lvert x_{iA}-x_{iB}\rvert}{\sum_{i=1}^{N}\max(x_{iA},x_{iB})}$ Range: 0 to 1	$D_{AB} = \frac{a+b-2c}{a+b-c}$ Range: 0 to 1

- Similarity (**S**) or distance (**D**) coefficients in common use for similarity searching in chemical databases.
- For binary data:
  - **a** is defined as the number of bits set to “1” in molecule **A**
  - **b** as the number of bits set to “1” in molecule **B**
  - **c** as the number of bits that are “1” in both **A** and **B**.

[Willett et al. 1998.]



Basic bit count terms of similarity

Symbol	Description	
$onlyA\|$	number of bits set "on" in fingerprint A but not in B	
$onlyB\|$	number of bits set "on" in fingerprint B but not in A	
$bothAB\|$	number of bits set "on" in both fingerprints	
$neitherAB\|$	number of bits set "off" in both fingerprints	
$\|A\|$	number of bits set "on" in fingerprint A	
$\|B\|$	number of bits set "on" in fingerprint B	

Euclidean

Formula:

$$Sim_{Euclid}(A,B) = \sqrt{\frac{bothAB+neitherAB}{onlyA+onlyB+bothAB+neitherAB}}$$

Range:

$[0.0 - 1.0]$

p = A
q = B

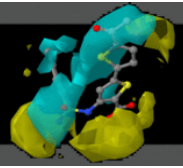
$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1-p_1)^2+(q_2-p_2)^2+\cdots+(q_n-p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}.$$

Example:

A	1	0	1	1	1	0	0	0
B	1	1	0	1	1	0	1	0

$$\sqrt{\frac{bothAB+neitherAB}{onlyA+onlyB+bothAB+neitherAB}} = \sqrt{\frac{3+1}{1+2+3+1}} = \sqrt{\frac{4}{8}} = 0.707$$



Manhattan

Formula:

$$Sim_{Manhattan}(A,B) = \frac{onlyA+onlyB}{onlyA+onlyB+bothAB+neitherAB}$$

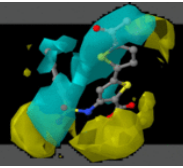
Range:

$$[1.0 - 0.0]$$

Example:

A	1	0	1	1	1	0	0	0
B	1	1	0	1	1	0	1	0

$$\frac{onlyA+onlyB}{onlyA+onlyB+bothAB+neitherAB} = \frac{1+2}{1+2+3+2} = \frac{3}{8} = 0.375$$



Cosine

Formula:

$$Sim_{Cosine}(A,B) = \frac{bothAB}{\sqrt{(onlyA+bothAB)*(onlyB+bothAB)}}$$

Range:

$[0.0 - 1.0]$

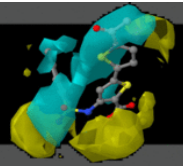
$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Example:

A	1	0	1	1	1	0	0	0
B	1	1	0	1	1	0	1	0

$$\frac{bothAB}{\sqrt{(onlyA+bothAB)*(onlyB+bothAB)}} = \frac{3}{\sqrt{(1+3)*(2+3)}} = \frac{3}{\sqrt{20}} = 0.67$$

Calculates the ratio of the bits in common to the geometric mean of the number of "on" bits in the two fingerprints.



Tanimoto (Jaccard)

Formula:

$$Sim_{Tanimoto}(A,B) = \frac{bothAB}{|A|+|B|-bothAB} = \frac{bothAB}{onlyA+onlyB+bothAB}$$

Range:

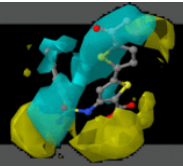
$$[0.0 - 1.0]$$

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

Example:

A	1	0	1	1	1	0	0	0
B	1	1	0	1	1	0	1	0

$$\frac{bothAB}{onlyA+onlyB+bothAB} = \frac{3}{1+2+3} = \frac{3}{6} = 0.5$$



Tversky

Formula:

$$Sim_{Tversky}(A,B) = \frac{bothAB}{\alpha * onlyA + \beta * onlyB + bothAB}$$

$$S(X,Y) = \frac{|X \cap Y|}{|X \cap Y| + \alpha|X - Y| + \beta|Y - X|}$$

The Tversky similarity measure is asymmetric. Setting the parameters $\alpha = \beta = 1.0$ is identical to using the *Tanimoto* measure.

The factor α weights the contribution of the first 'reference' molecule. The larger α becomes, the more weight is put on the bit setting of the reference molecule.

Setting α = β = 1 produces the *Tanimoto coefficient*.
Setting α = β = 0.5 produces *Dice's coefficient.*

Range:

variable

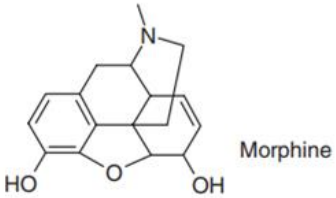
If we consider **A** to be the prototype and **B** to be the variant, then **α** corresponds to the weight of the prototype and **β** corresponds to the weight of the variant.

Example:

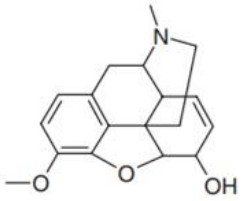
A	1	0	1	1	1	0	0	0
B	1	1	0	1	1	0	1	0

$$\frac{bothAB}{\alpha * onlyA + \beta * onlyB + bothAB}(\alpha = 2.0, \beta = 1.0) = \frac{3}{2.0*1 + 1.0*2 + 3} = \frac{3}{7} = 0.438$$

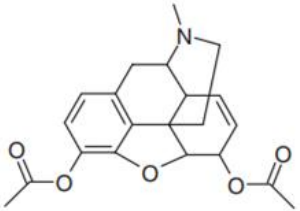
2D similarity (1)



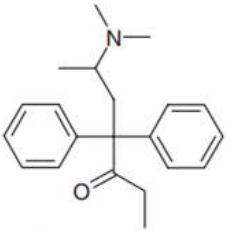
Morphine



Codeine
0.99 similar



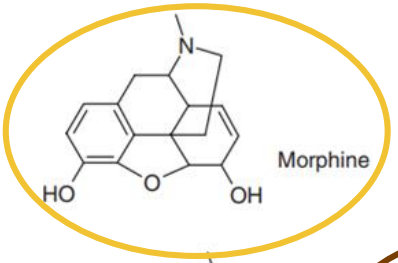
Heroin
0.95 similar



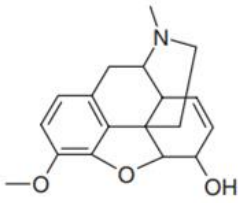
Methadone
0.20 similar

Similarities to morphine calculated using Daylight fingerprints and the Tanimoto coefficient.

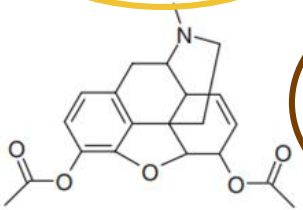
2D similarity (2)



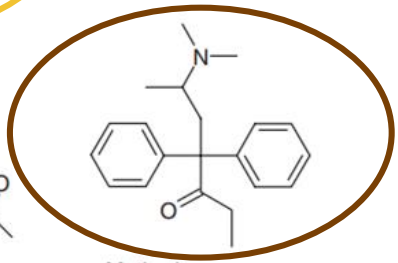
Molecular recognition depends on the 3D structure and properties (e.g. electrostatics and shape) of a molecule rather than the underlying substructure(s).



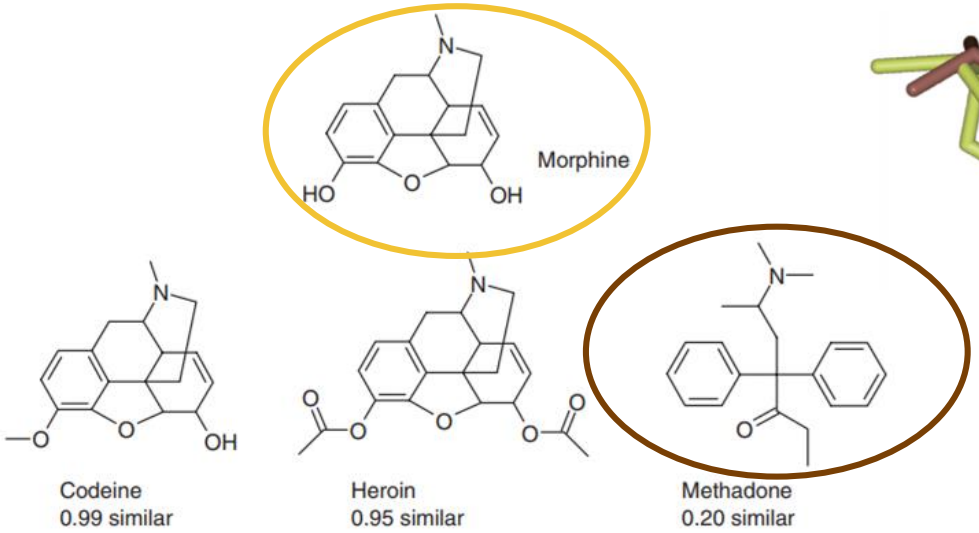
Codeine
0.99 similar

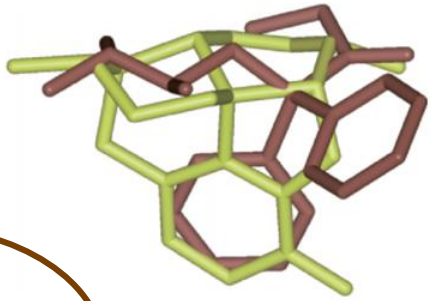


Heroin
0.95 similar

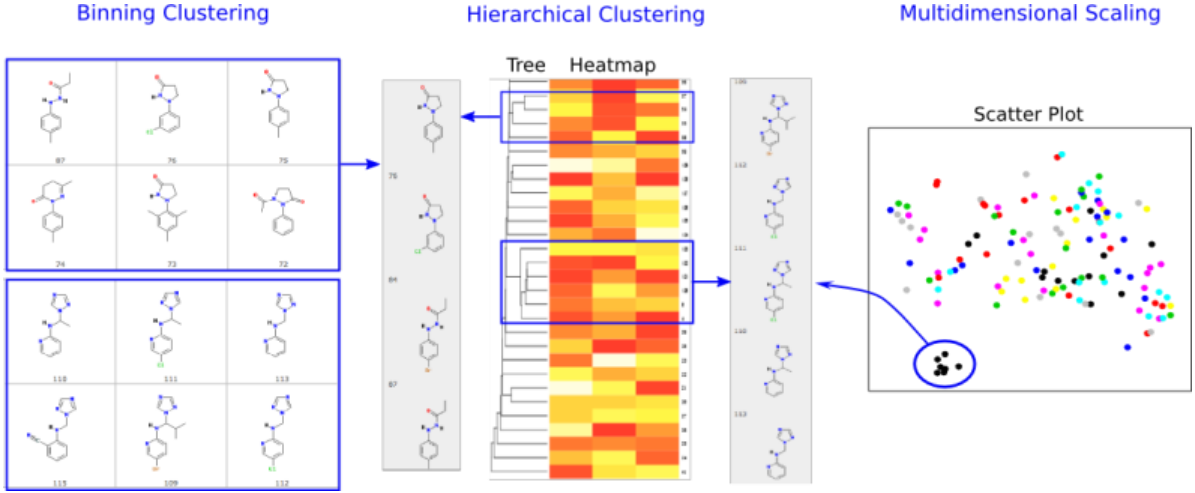


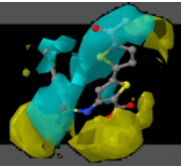
Methadone
0.20 similar


N
HO
O
OH
Morphine
N
O
O
OH
Codeine
0.99 similar
N
O
O
O
O
Heroin
0.95 similar
N
O
Methadone
0.20 similar



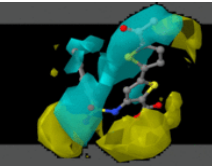
Cluster and diversity analysis



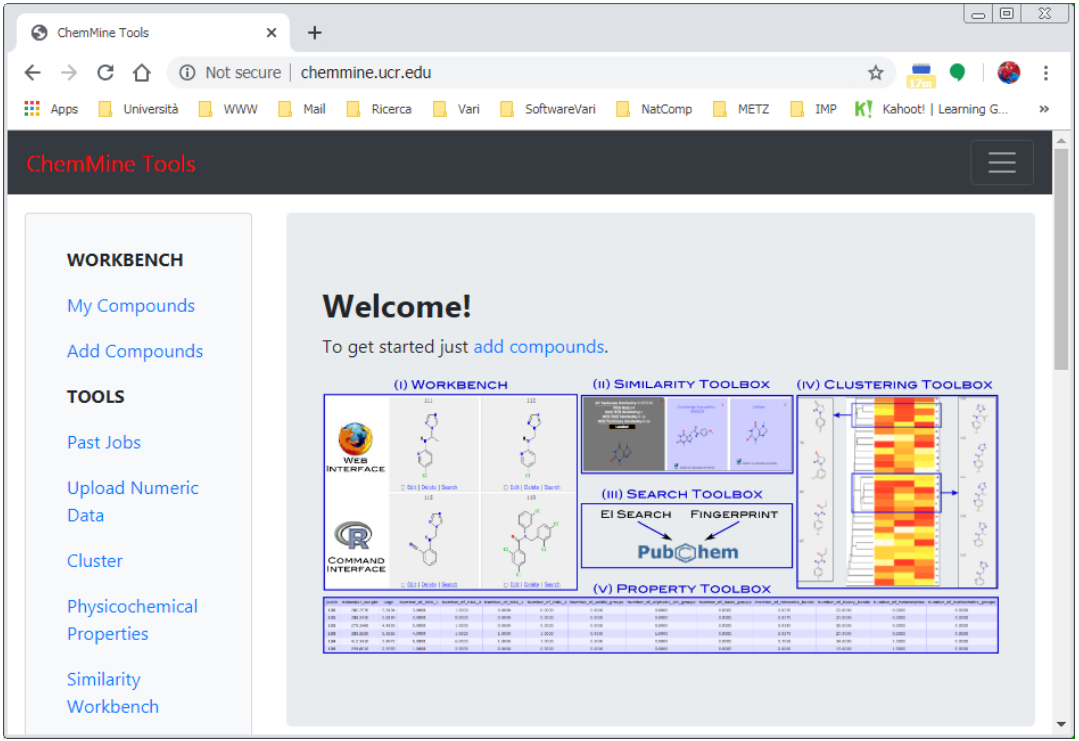


References

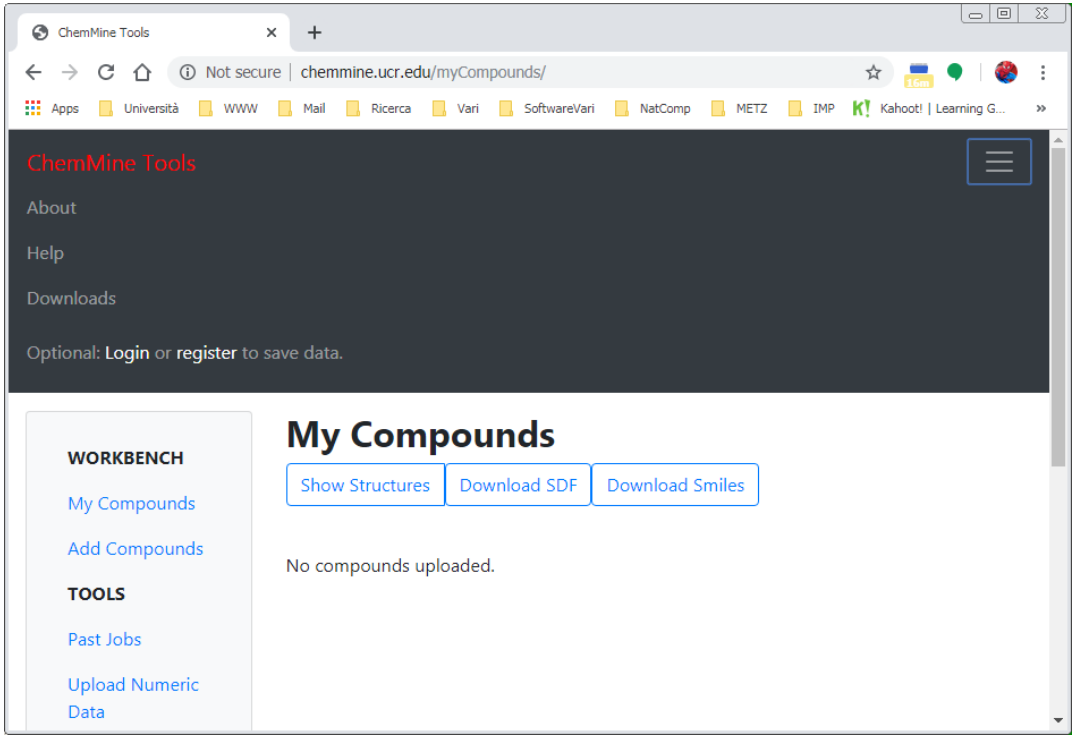
- Leach, Andrew R., and Valerie J. Gillet. An introduction to chemoinformatics. Springer Science & Business Media, 2007.
- Baldi, Pierre, and Ramzi Nasr. "When is chemical similarity significant? The statistical distribution of chemical similarity scores and its extreme values." Journal of chemical information and modeling 50.7 (2010): 1205-1222.
- Riniker, Sereina, and Gregory A. Landrum. "Open-source platform to benchmark fingerprints for ligand-based virtual screening." Journal of cheminformatics 5.1 (2013): 1.
- Maggiora, Gerald, et al. "Molecular Similarity in Medicinal Chemistry: Miniperspective." Journal of medicinal chemistry 57.8 (2013): 3186-3204.
- Bajusz, Dávid, Anita Rácz, and Károly Héberger. "Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?." Journal of cheminformatics 7.1 (2015): 1.
- Duesbury, Edmund, John Holliday, and Peter Willett. "Maximum common substructure-based data fusion in similarity searching." Journal of chemical information and modeling 55.2 (2015): 222-230.
- Award Symposium 2015: Jürgen Bajorath. American Chemical Society, 2016.
- Schyman, Patric, Ruifeng Liu, and Anders Wallqvist. "General Purpose 2D and 3D Similarity Approach to Identify hERG Blockers." Journal of chemical information and modeling 56.1 (2016): 213-222.
- http://www.daylight.com/dayhtml/doc/theory/theory.finger.html
- https://docs.eyesopen.com/toolkits/python/graphsimtk/measure.html



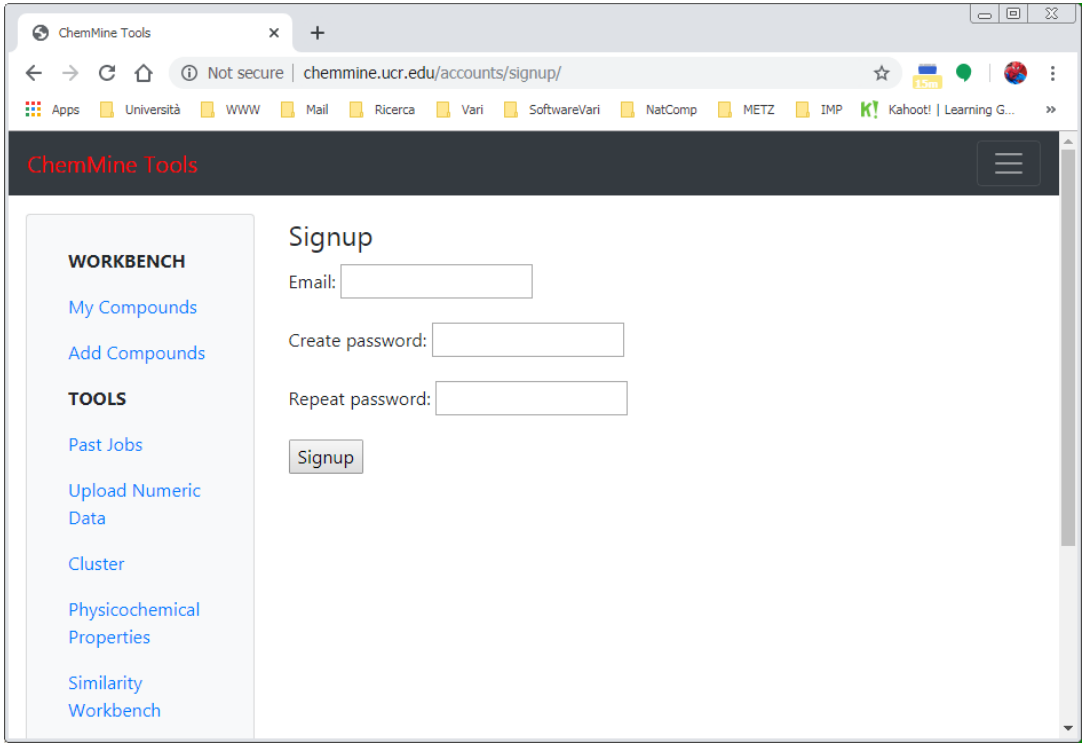
ChemMine Tools



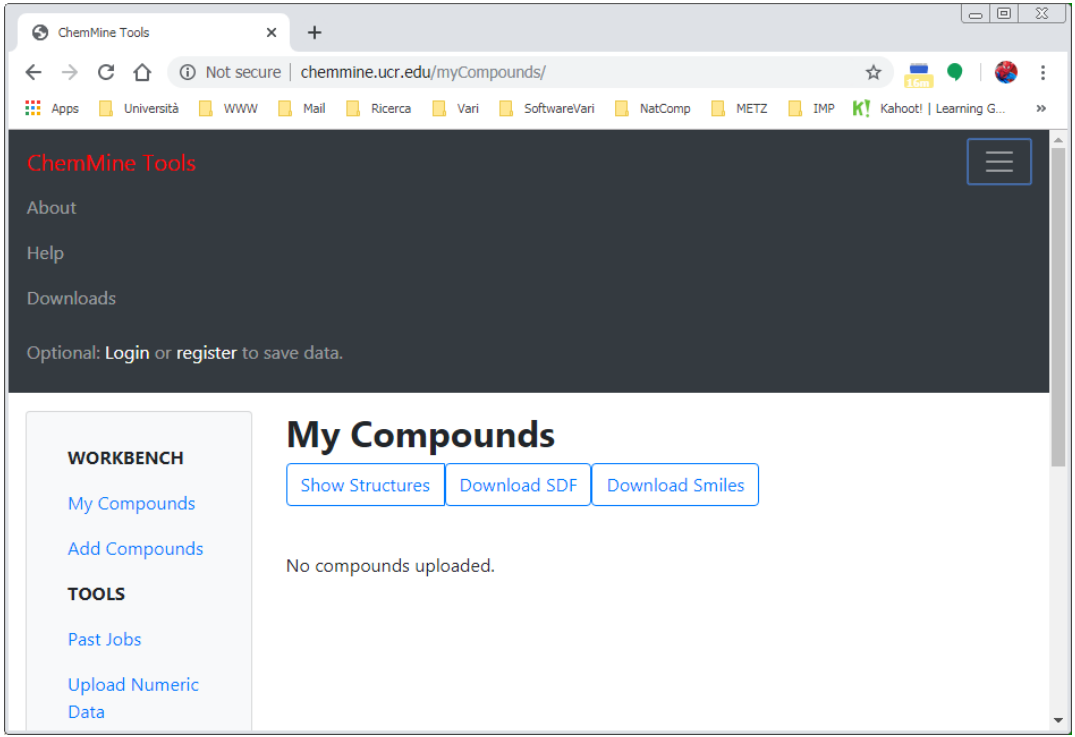
ChemMine Tools



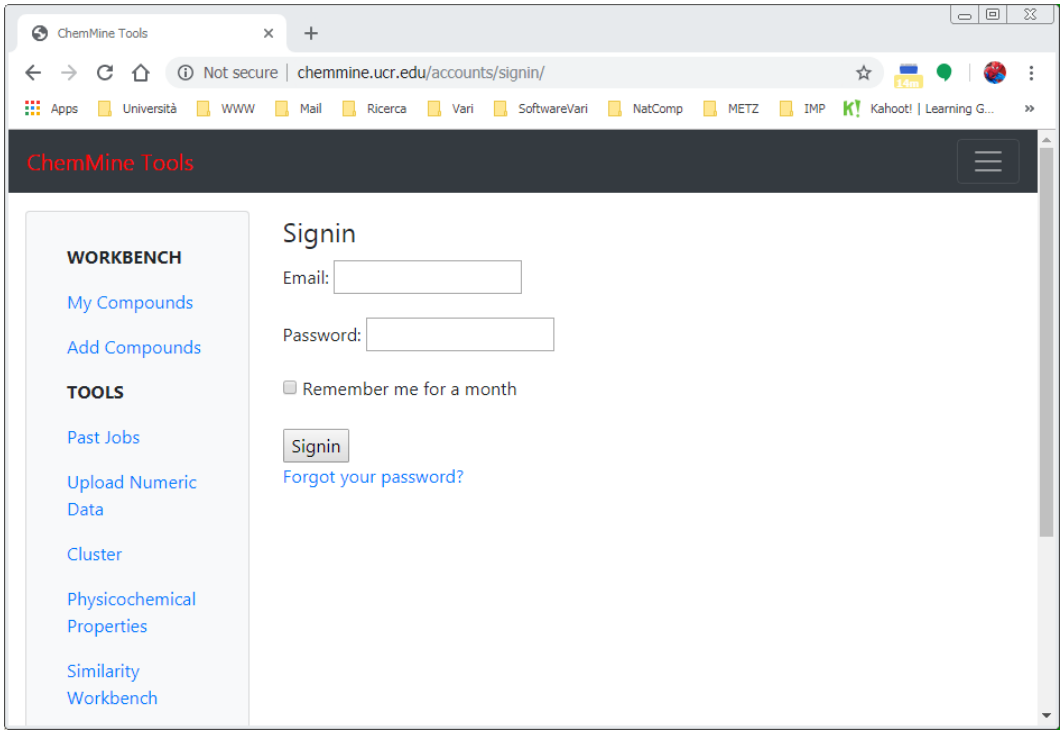
ChemMine Tools

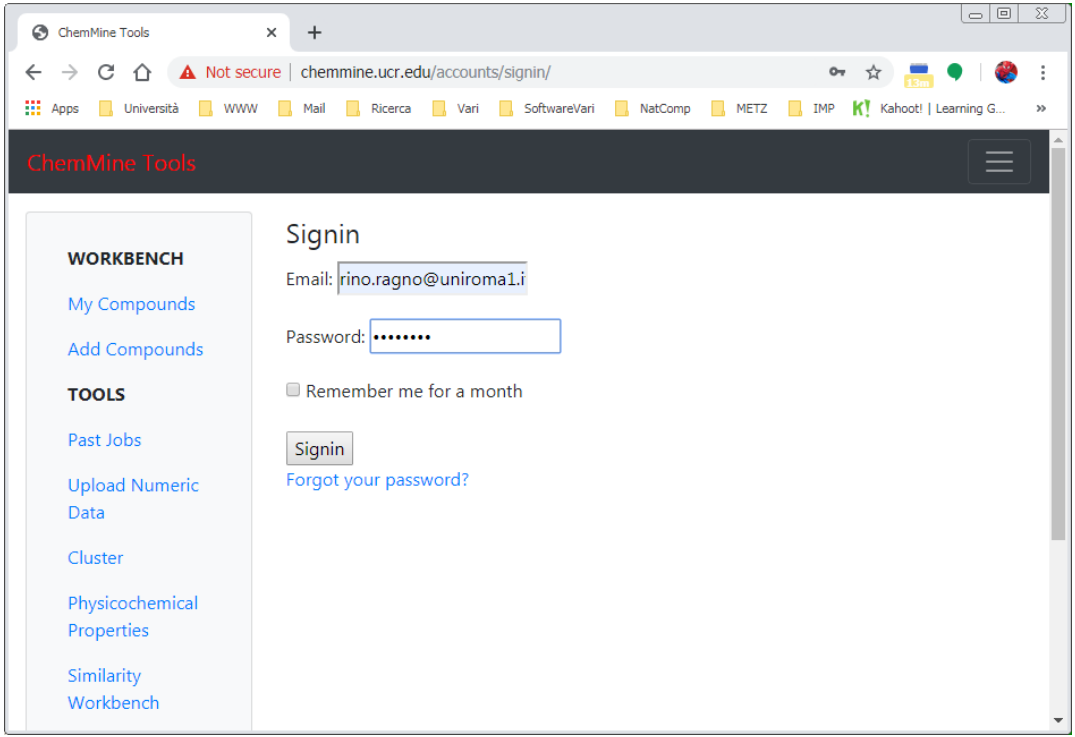


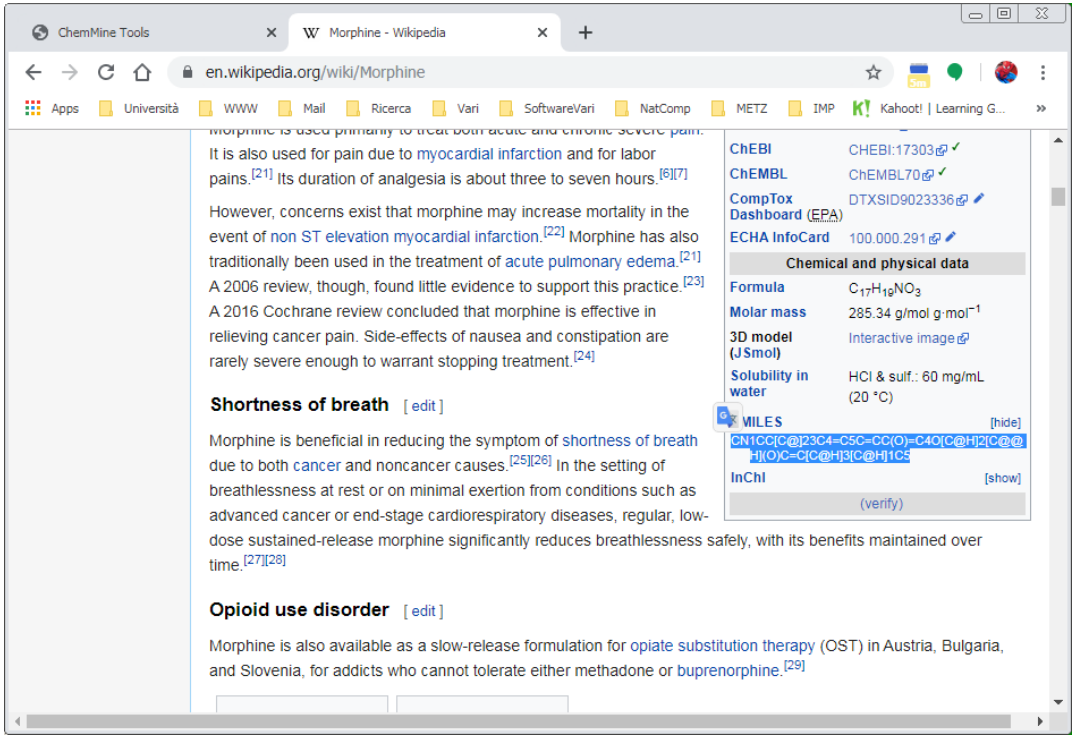
ChemMine Tools

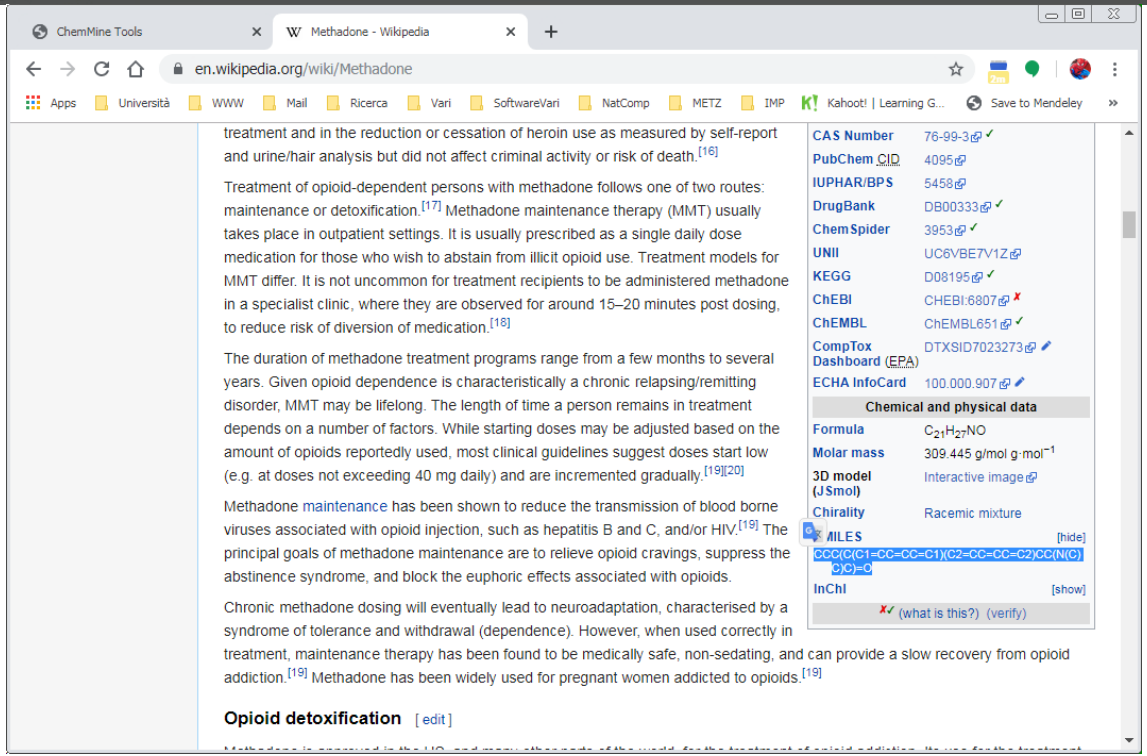


ChemMine Tools

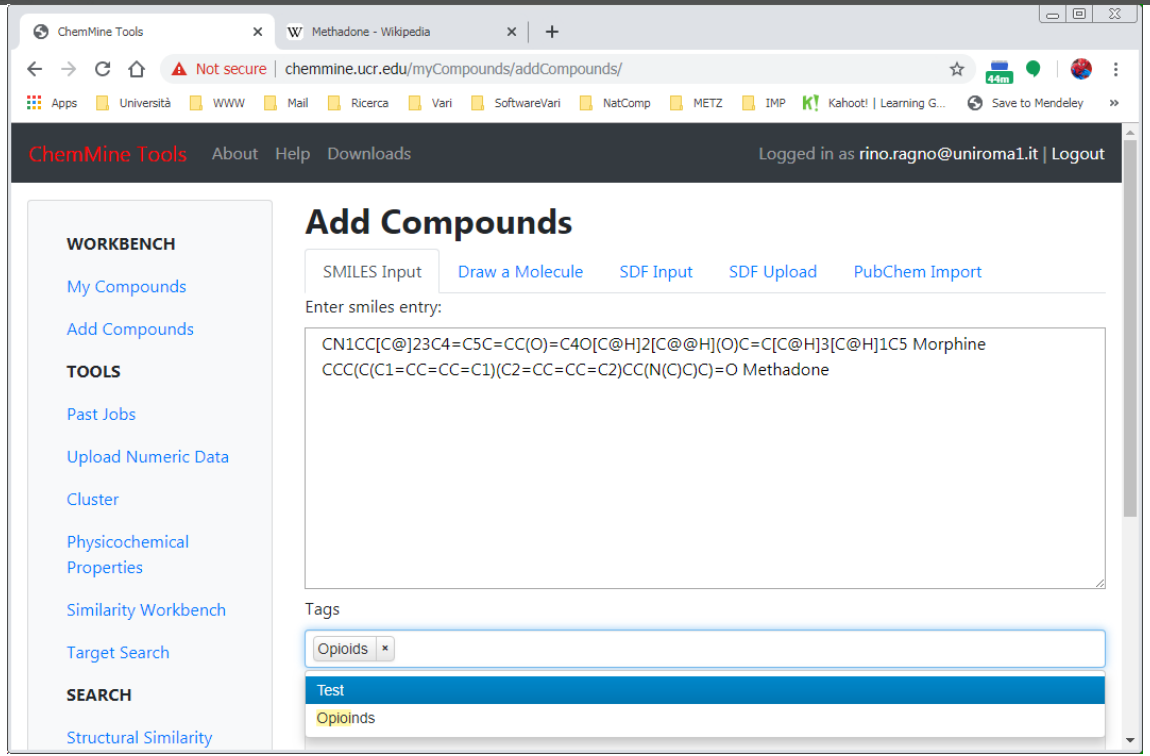


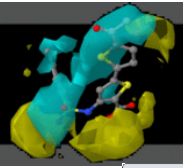

ChemMine Tools
Not secure
chemmine.ucr.edu/accounts/signin/
Apps
Università
WWW
Mail
Ricerca
Vari
SoftwareVari
NatComp
METZ
IMP
Kahoot! | Learning G...
ChemMine Tools
WORKBENCH
My Compounds
Add Compounds
TOOLS
Past Jobs
Upload Numeric Data
Cluster
Physicochemical Properties
Similarity Workbench
Signin
Email: rino.ragno@uniroma1.i
Password:
Remember me for a month
Signin
Forgot your password?


ChemMine Tools
Morphine - Wikipedia
en.wikipedia.org/wiki/Morphine
Apps
Università
WWW
Mail
Ricerca
Vari
SoftwareVari
NatComp
METZ
IMP
Kahoot! | Learning G...
It is also used for pain due to myocardial infarction and for labor pains.[21] Its duration of analgesia is about three to seven hours.[6][7]
However, concerns exist that morphine may increase mortality in the event of non ST elevation myocardial infarction.[22] Morphine has also traditionally been used in the treatment of acute pulmonary edema.[21] A 2006 review, though, found little evidence to support this practice.[23] A 2016 Cochrane review concluded that morphine is effective in relieving cancer pain. Side-effects of nausea and constipation are rarely severe enough to warrant stopping treatment.[24]
Shortness of breath [edit]
Morphine is beneficial in reducing the symptom of shortness of breath due to both cancer and noncancer causes.[25][26] In the setting of breathlessness at rest or on minimal exertion from conditions such as advanced cancer or end-stage cardiorespiratory diseases, regular, low-dose sustained-release morphine significantly reduces breathlessness safely, with its benefits maintained over time.[27][28]
Opioid use disorder [edit]
Morphine is also available as a slow-release formulation for opiate substitution therapy (OST) in Austria, Bulgaria, and Slovenia, for addicts who cannot tolerate either methadone or buprenorphine.[29]
ChEBI CHEBI:17303
ChEMBL ChEMBL70
CompTox Dashboard (EPA) DTXSID9023336
ECHA InfoCard 100.000.291
Chemical and physical data
Formula $C_{17}H_{19}NO_3$
Molar mass 285.34 g·mol^{-1}
3D model (JSmol) Interactive image
Solubility in water HCl & sulf.: 60 mg/mL (20 °C)
MILES [hide]
CN1CC[C@]23C4=C5C=CC(O)=C4O[C@H]2[C@@H](O)C=C[C@H]3[C@H]1C5
InChI [show]
(verify)


ChemMine Tools
Methadone - Wikipedia
en.wikipedia.org/wiki/Methadone
Apps
Università
WWW
Mail
Ricerca
Vari
SoftwareVari
NatComp
METZ
IMP
Kahoot! | Learning G...
Save to Mendeley
treatment and in the reduction or cessation of heroin use as measured by self-report and urine/hair analysis but did not affect criminal activity or risk of death.[16]
Treatment of opioid-dependent persons with methadone follows one of two routes: maintenance or detoxification.[17] Methadone maintenance therapy (MMT) usually takes place in outpatient settings. It is usually prescribed as a single daily dose medication for those who wish to abstain from illicit opioid use. Treatment models for MMT differ. It is not uncommon for treatment recipients to be administered methadone in a specialist clinic, where they are observed for around 15–20 minutes post dosing, to reduce risk of diversion of medication.[18]
The duration of methadone treatment programs range from a few months to several years. Given opioid dependence is characteristically a chronic relapsing/remitting disorder, MMT may be lifelong. The length of time a person remains in treatment depends on a number of factors. While starting doses may be adjusted based on the amount of opioids reportedly used, most clinical guidelines suggest doses start low (e.g. at doses not exceeding 40 mg daily) and are incremented gradually.[19][20]
Methadone maintenance has been shown to reduce the transmission of blood borne viruses associated with opioid injection, such as hepatitis B and C, and/or HIV.[19] The principal goals of methadone maintenance are to relieve opioid cravings, suppress the abstinence syndrome, and block the euphoric effects associated with opioids.
Chronic methadone dosing will eventually lead to neuroadaptation, characterised by a syndrome of tolerance and withdrawal (dependence). However, when used correctly in treatment, maintenance therapy has been found to be medically safe, non-sedating, and can provide a slow recovery from opioid addiction.[19] Methadone has been widely used for pregnant women addicted to opioids.[19]
Opioid detoxification [edit]
CAS Number 76-99-3
PubChem CID 4095
IUPHAR/BPS 5458
DrugBank DB00333
ChemSpider 3953
UNII UC6VBE7V1Z
KEGG D08195
ChEBI CHEBI:6807
ChEMBL ChEMBL651
CompTox Dashboard (EPA) DTXSID7023273
ECHA InfoCard 100.000.907
Chemical and physical data
Formula C21H27NO
Molar mass 309.445 g·mol−1
3D model (JSmol) Interactive image
Chirality Racemic mixture
SMILES [hide]
CCC(C(C1=CC=CC=C1)C2=CC=CC=C2)CC(N(C)C)=O
InChI [show]
(what is this?) (verify)

ChemMine Tools





ChemMine Tools

by www.RCMD.it

ChemMine Tools | Methadone - Wikipedia

Not secure | chemmine.ucr.edu/myCompounds/

Apps | Università | WWW | Mail | Ricerca | Vari | SoftwareVari | NatComp | METZ | IMP | Kahoot! | Learning G... | Save to Mendeley

ChemMine Tools About Help Downloads Logged in as rino.ragno@uniroma1.it | Logout

WORKBENCH

My Compounds

Add Compounds

TOOLS

Past Jobs

Upload Numeric Data

Cluster

Physicochemical Properties

Similarity Workbench

Target Search

SEARCH

Structural Similarity

Success: Added 2 compounds. ×

My Compounds

Show Structures | Download SDF | Download Smiles

Select all | Deselect all | Add Tags | Remove Tags | Delete Selected

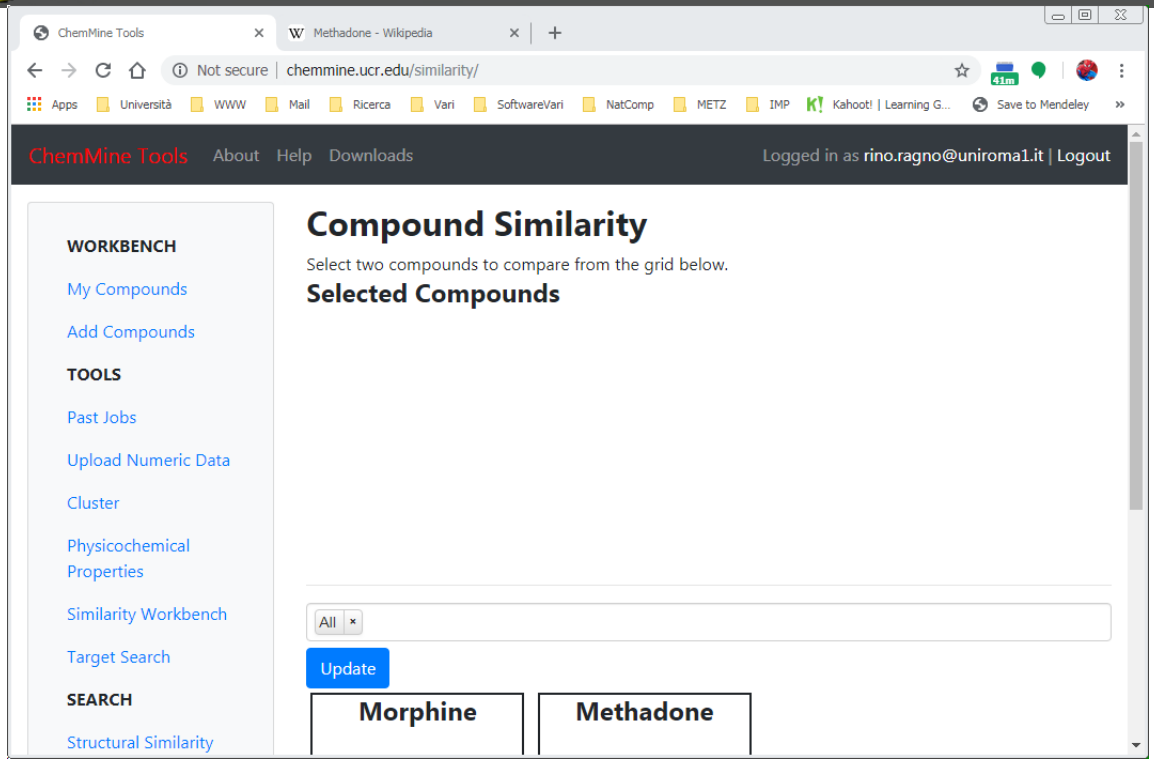
Showing 1 to 2 of 2 entries

Search:

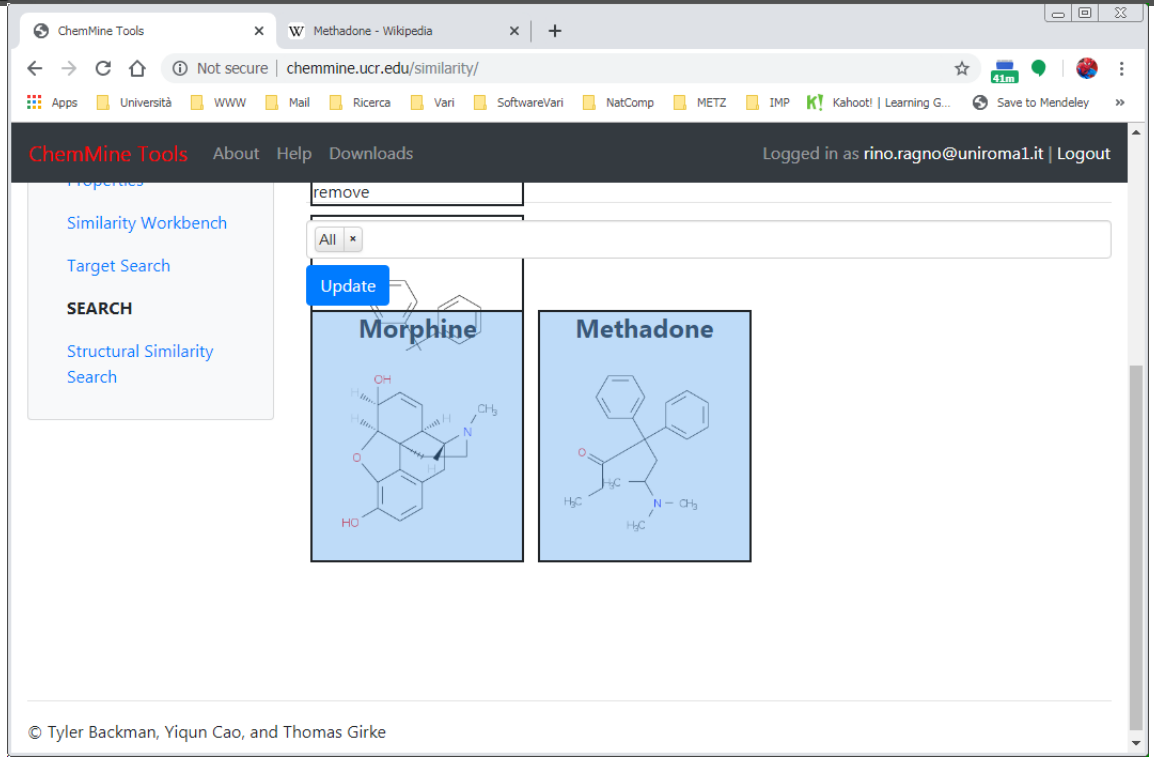
Show 10 entries

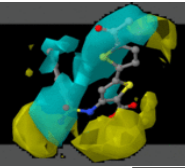
CID	Name	Formula	Tags	Options
Methadone		C21H27NO	Test, Opioinds, Opioids,	
Morphine		C17H19NO3	Test, Opioinds, Opioids,	

ChemMine Tools



ChemMine Tools





ChemMine Tools

by www.RCMD.it

ChemMine Tools | Methadone - Wikipedia

Not secure | chemmine.ucr.edu/similarity/

Apps | Università | WWW | Mail | Ricerca | Vari | SoftwareVari | NatComp | METZ | IMP | Kahoot! | Learning G... | Save to Mendeley

ChemMine Tools About Help Downloads Logged in as rino.ragno@uniroma1.it | Logout

WORKBENCH

My Compounds

Add Compounds

TOOLS

Past Jobs

Upload Numeric Data

Cluster

Physicochemical Properties

Similarity Workbench

Target Search

SEARCH

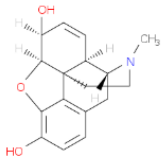
Structural Similarity

Compound Similarity

Select two compounds to compare from the grid below.

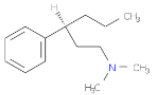
Selected Compounds

Morphine



remove

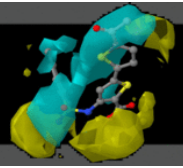
AP Tanimoto: 0.202597
MCS Tanimoto: 0.5172
MCS Size: 15
MCS Min: 0.7143
MCS Max: 0.6522
SMILES: [C@@H](c1ccccc1)(CCN(C)C)CCC
Morphine



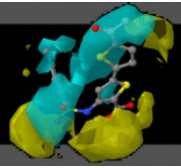
All ×

Update

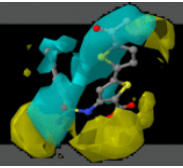
Morphine **Methadone**



ChemMine Tools



ChemMine Tools



ChemMine Tools